Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0

10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0

Zack Hodari, Catherine Lai, Simon King

Centre for Speech Technology Research, University of Edinburgh, UK
{zack.hodari, c.lai, Simon.King}@ed.ac.uk

Abstract

In English, prosody adds a broad range of information to segment sequences, from information structure (e.g. contrast) to stylistic variation (e.g. expression of emotion). However, when learning to control prosody in text-to-speech voices, it is not clear what exactly the control is modifying. Existing research on discrete representation learning for prosody has demonstrated high naturalness, but no analysis has been performed on what these representations capture, or if they can generate meaningfully-distinct variants of an utterance. We present a phrase-level variational autoencoder with a multi-modal prior, using the mode centres as '*intonation codes*'. Our evaluation establishes which *intonation codes* are perceptually distinct, finding that the *intonation codes* from our multi-modal latent model were significantly more distinct than a baseline using k-means clustering. We carry out a follow-up qualitative study to determine what information the *codes* are carrying. Most commonly, listeners commented on the *intonation codes* having a statement or question style. However, many other affect-related styles were also reported, including: emotional, uncertain, surprised, sarcastic, passive aggressive, and upset. Finally, we lay out several methodological issues for evaluating distinct prosodies.
Index Terms: speech synthesis, intonation modelling, prosodic variation, speech perception, discrete representation learning, variational autoencoder

1. Introduction

In text-to-speech synthesis (TTS), the natural variability of prosody is often not accounted for. Current TTS systems default to the production of average prosody [1]: monotonous and boring speech. Synthetic voices do not take contextual variation into account during training, thus different prosodies are seen as noise and only the mean is learnt. This sort of overly smoothed speech can be fatiguing to listen to in long form speech. However, relevant context can be very wide ranging and much of this can be expensive or impractical to obtain. For example, previous work has identified consistent variations in prosody with respect to structural elements in the discourse context [2, 3, 4] and sentence level information structure [5, 6]. Variation has also been identified with respect to specific speaker attitudes [7, 8, 9] and stances [10, 11, 12].

Given this, a successful TTS system should be able to produce a large variety of plausible prosodic forms for a given utterance. However, current TTS systems often rely on pre-specified linguistic context features to guide prosodic realisations of synthetic speech. Unsurprisingly, annotated speech data with wide coverage of suitably-rich contextual information is not widely available. So, in order to develop TTS models that generate plausible and appropriate prosody given a specific context, we propose to split the problem in half: *controllability*—designing a system that can produce distinct renditions of isolated sentences; and *appropriateness*—choosing the most appropriate rendition using contextual information. This paper presents work on the first task: learning a prosodic representation capable of producing distinct renditions of a single sentence. Importantly, we do not attempt to identify the most appropriate or most likely prosody given a pre-specified context. Instead, our goal is to verify that different renditions produced by our representation are *perceived* as distinct, and whether they convey different information or intent.

Most TTS research on controllability focuses on emotion or emphasis [13, 14]. Conversely, more fundamental prosody research has focused on how acoustic-phonetic features map to linguistic categories. We want to bridge this gap in order to make advances in both together, by determining what meaning listeners perceive in renditions from controllable TTS. Recent phonetic studies support the idea that both categorical and continuous features are integral to prosodic variation [15, 16]. In line with this, we learn *discrete* representations which can potentially capture categorical differences often associated with phrasing and prominence, but also allow for the generation of fine-grained phonetic differences, which vary the perception of expressivity, emphasis, and speaker affect.

We use a variational autoencoder with a multi-modal prior (Section 3.3) to learn a discrete representation of F0. We evaluate what these '*intonation codes*' capture through subjective (Section 6.1) and qualitative (Section 6.2) tests.

2. Related work

Controllable TTS has been approached from both supervised and unsupervised perspectives. Henter et al. [14] demonstrated that both can achieve the same quality for emotion control.

Unsupervised representation learning in TTS typically uses a continuous representation (i.e. $\mathbb{R}^n$) at the sentence level [17, 18, 1], but this becomes increasingly difficult to interpret for $n \gtrsim 3$. Poor interpretability limits the range of practical use cases. For example, [19, 18] are limited to transferring style from another natural utterance. To address practical limitations, high-dimensional representations can be predicted automatically, perhaps using the current utterance [20, 21]. Discrete representations are another way to address interpretability [22], and can also be paired with automatic prediction from text [23].

Prosody should be modelled in the correct *domain*. While most approaches [17, 19, 14, 18, 1] operate on sentences, the sentence domain may not be the most appropriate for a fixed-sized prosodic representation. For example, sentences contain a variable number of prosodic phrases. It is likely that by working in the sentence domain, something closer to sentence-style, as opposed to prosody, is captured. Much less work has been done on prosodically-appropriate domains. Wang et al. [23] compare a discrete representation of F0 in the phrase domain to smaller and longer domains. Reconstruction performance clearly shows

10.21437/SpeechProsody.2020-197

that these fixed-sized representations are less accurate for longer domains which can contain more information.

Although claims of expressivity or prosody control are often made, variability or controllability are often not evaluated. In [18], prosody reconstruction measures the model's topline performance, and prosody transfer is demonstrated qualitatively, but interpreting the latent space, or choosing the best rendition was not tackled. Tyagi et al. [21] present a unit selection-like system for prosody generation. Prosody embeddings of the training sentences act as templates and are chosen using a linguistically-informed target cost and an acoustic join cost. They evaluate general appropriateness of isolated sentences using linguistic expert listeners. However, a single best rendition is predicted without reference to additional context.

Without sufficient context, appropriateness is arbitrary. An isolated sentence has multiple valid prosodies with varying frequency of occurrence. Two approaches to determining appropriateness would be: rating appropriateness given a specific context; or collecting contexts that make a given prosody likely (cf. Section 6.2 (iii)). However, previous studies highlight how listener perception of prosody can be affected by both context and what listeners are told to attend to [24, 25]. Thus, our current work focuses on generating distinct prosodic renditions and exploring how differences in prosody are perceived, deferring full-scale appropriateness evaluation for future work.

3. Learning a discrete prosodic representation

We present two methods for learning '*intonation codes*': a baseline using an autoencoder (AE) and k-means in Section 3.2; and our proposed method using a variational autoencoder (VAE) with learned multi-modal structure in the latent space in Section 3.3. However, first we address the issue of domain.

3.1. Prosodic phrasing

An obvious domain for prosodic control is the prosodic phrase, but accurately locating prosodic phrase boundaries (breaks) requires manual annotation. While there is a correlation between syntactic and prosodic *structure* [26], mismatches between syntactic and prosodic phrase boundaries are common [27]. So, instead we adopt Liberman and Church's notion of *chinks 'n chunks* [28] which aims to identify contiguous units of text that map more appropriately to phrases for TTS.[1]

Chinks 'n chunks is a simple heuristic parser that takes advantage of the right-branching nature of English; content words tend to occur towards the end of phrases and function words towards the beginning. However, since certain word types can behave like either, Liberman and Church define two categories:

chink	—	function words	+	tensed verbs
chunk	—	content words	+	objective pronouns

Tensed verbs can behave like auxiliaries, thus starting a phrase. Objective pronouns can behave like nouns, thus acting as content words. The parsing algorithm is a simple greedy match of {**chink* chunk***}, see Table 1 for phrase examples (in bold).

3.2. Baseline: two-stage clustering

Our baseline (top of Figure 1) has two stages: learn continuous embeddings z using an AE; cluster the training data embeddings using k-means. We call the clusters z_q 'intonation codes'.

[1] We thank Oliver Watts for suggesting this method, and helping with the finer details of the parser.

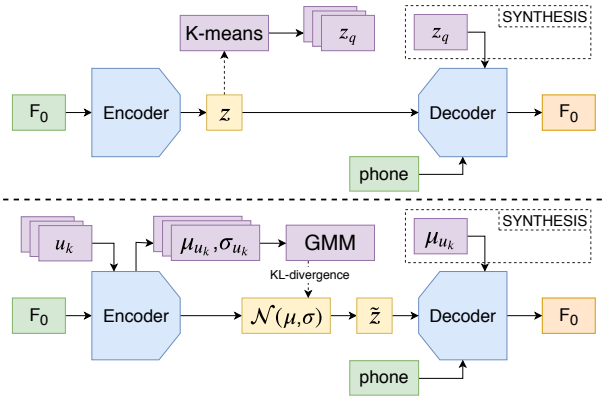

Figure 1: *Architecture for* $\text{AE}_{\text{K-MEANS}}$ *(top) and* VAE_{VAMP} *(bottom). Yellow and purple indicate phrase-level portions, while purple shows specifically where discreteness is added.*

Note that an AE's reconstruction loss (indirectly) encourages similar inputs to locate close to each other in the embedding space, thus imposing an implicit distance (without scale). For this reason, unsupervised clustering is feasible, though the two-stage process may lead to sub-optimal *intonation codes*.

3.3. Probabilistic multi-modal latent space

The two-stage approach in Section 3.2 is limiting as the AE will not necessarily structure its space into clusters. Ideally, the embedding space and clustered structure would be learnt jointly.

VAEs [29] are subject to a prior reflecting our assumptions about the underlying latent factors that describe the data. This prior directly enforces distance and scale on the space, which can in turn be used to enforce a clustered structure. In [29], a unimodal Gaussian is used, but we want to find distinct prosodic behaviours in our data. Hence, we use a variational mixture of posteriors (VAMP) prior [30] (bottom of Figure 1, in purple).

In simple terms, the VAMP prior is a Gaussian mixture model (GMM), whose parameters are learned jointly with the rest of the model. However, we do not learn the GMM parameters directly, instead we learn K '*pseudo-inputs*' $\{\mathbf{u}_k\}_{k=1}^{K}$, where K is a hyperparameter. These *pseudo-inputs* are not real inputs, they are parameters learned through backpropagation. Each GMM component is given by a *pseudo-input*'s approximate posterior $p(\mathbf{z}_k \mid \mathbf{u}_k) = \mathcal{N}(\mathbf{z}_k; \boldsymbol{\mu}_{\mathbf{u}_k}, \boldsymbol{\sigma}_{\mathbf{u}_k})$. Here, we define our *intonation codes* using GMM component centres $\boldsymbol{\mu}_{\mathbf{u}_k}$.

Since we learn *pseudo-inputs* and not GMM parameters to define our prior, we are learning parameters in the input space. Tomczak and Welling [30] demonstrated this for fixed-size images; we present what we believe to be the first application of VAMP to sequence data (F0 contours). Therefore, we have to contend with learning a sequence of parameters for each *pseudo-input*. While it may be possible to learn the sequence lengths, in this work we choose to fix the number of frames of each *pseudo-input* at initialisation. See Section 5 for more discussion on *pseudo-input* sequence length.

4. Data

Our choice of training data is motivated by the need for interesting variation: if the data is very stylistically consistent, there will be too little variation to capture through *intonation codes*. We therefore use the Blizzard Challenge 2018 dataset [31] consisting of stories read in an expressive style for a 4–6 year old audience. In total it contains 6.5 hours (~7,250 sentences) of

professionally-recorded speech from a female speaker of standard southern British English. Three stories were held out for the listening test: Goldilocks and the Three Bears, The Boy Who Cried Wolf, and The Enormous Turnip.

While this is not conversational data, it does contain character voices and direct speech. Our *intonation codes* may capture the child audiobook style as opposed to prosody typically seen in dialogue. However, this work serves as a proof of concept that we will later validate using dialogue data [32, 33].

5. System details

Our two models,[2] $\text{AE}_{\text{K–MEANS}}$ and VAE_{VAMP}, both have an auto-encoder structure, encoding and reconstructing mean-variance normalised logF0, delta, and delta-delta features. MLPG [34] is used for F0 generation using global standard deviation. This F0 contour is then synthesised with natural spectral features using WORLD [35] and a frame-shift of 5ms. For TTS the *intonation codes* for the decoder must be chosen without using natural F0, as discussed in Section 6.

The encoders and decoders for both systems are as follows: a feedforward layer with 256 units, followed by three recurrent layers using gated recurrent cells with 64 units. Finally, outputs are projected to the required dimension. Both decoders are conditioned on one-hot phone identity. We found that a full linguistic specification limited the range of variation captured in F0. Phone identity was upsampled to frame-level using forced alignment durations.[3] The encoders are clocked at the frame-level, so to get the sequence of phrase-level *intonation codes* for a sentence, we take the encoder outputs at the last frame of each phrase, and assign each output to a cluster/mode. The *intonation codes* are defined as follows:

$\text{AE}_{\text{K–MEANS}}$ $-\mathbf{z}_q$ (cluster centroids)

VAE_{VAMP} $-\boldsymbol{\mu}_{\mathbf{u}_k}$ (mean of *pseudo-input* approx. posteriors)

We use 20 clusters for $\text{AE}_{\text{K–MEANS}}$, and 20 *pseudo-inputs* for VAE_{VAMP}. As discussed earlier, we fix the sequence length of the *pseudo-inputs* at initialisation. Using the same sequence length for all *pseudo-inputs* was adequate and gave a stable model, if that sequence length is within the range seen in the training set: ~50 to ~500 frames. However, we obtained more distinct clusters by using varied *pseudo-input* sequence lengths. We used sequence lengths from 50 to 500 frames, with a step of 50 and repeating each length twice, for a total of 10 unique sequence lengths, and 20 *pseudo-inputs*. We used each sequence length twice to allow for multiple modes at each length.

Both models were trained for 100 epochs using Adam [39] with a learning rate increasing linearly from 0 to 0.005 over the first 8 epochs and then decaying proportional to the inverse square of the number of batches [40, Sec 5.3]. Our batch size is 32. The KL-divergence term in VAE_{VAMP} is weighted by zero during the first 5 epochs and increased linearly to 0.001 over 20 epochs. VAE_{VAMP} converged to a KL-divergence of 5.32. When using the oracle embedding, $\text{AE}_{\text{K–MEANS}}$ and VAE_{VAMP} achieved F0 RMSEs of 33.0Hz and 37.1Hz, respectively.

6. Evaluation

Recall that we aim to capture distinct prosodic characteristics using *intonation codes*, such as changes affecting information structure or strength of expressivity. Therefore the first step in evaluation is to determine whether changing the *intonation code* produces perceivable variation. Generating a new rendition of a sentence requires selecting a sequence of *intonation codes*: one per (*chink 'n chunk*-based) prosodic phrase. While both systems learn using multi-phrase sentences, we do not have a "language model" over these *codes*, as such we cannot know which *code* sequences are appropriate. Thus, we restrict the current work to sentences with one phrase and leave multi-phrase synthesis for future work. We randomly chose 12 single-phrase test sentences: 4 from each of the test set books in Table 1.

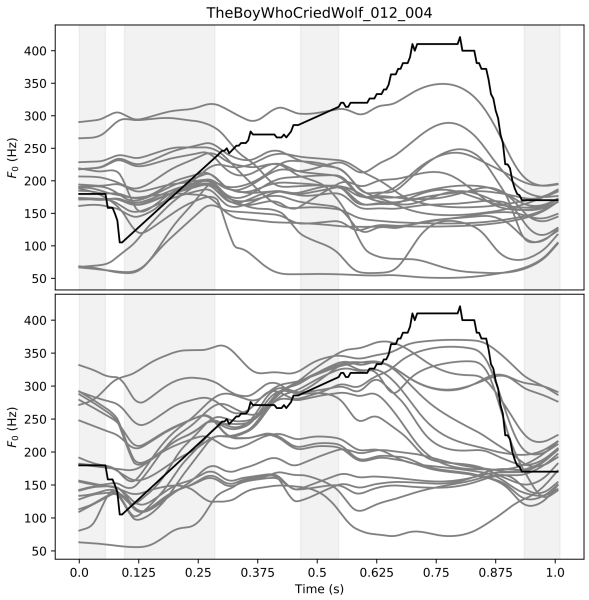

Figure 2: *20* codes *for* $\text{AE}_{\text{K–MEANS}}$ *(top) and* VAE_{VAMP} *(bottom) for the sentence: "What's the matter now?". The black line shows natural F0, interpolated linearly in unvoiced regions.*

6.1. Subjective evaluation

In a forced choice listening test, listeners were presented with two renditions of the same sentence and asked if they had "different intonation". We synthesised 40 renditions (20 $\text{AE}_{\text{K–MEANS}}$ clusters + 20 VAE_{VAMP} modes; Figure 2) of each of the 12 test sentences, from which we randomly chose 38 pairs. Each pair comprised two different renditions of the same sentence, both from the same system. A 2x2 Latin Square between-subjects design was used so that each listener heard all sentences, half the pairs from $\text{AE}_{\text{K–MEANS}}$ and half the pairs from VAE_{VAMP}. Across two listeners all pairs were presented once. 22 native English-speaking participants each took around 45 minutes to complete the test, for which they were paid £8.

Taking the results per system in a binomial significance test, we find that overall each system produced significant perceptual differences (Figure 3). The rate of perceptual difference for VAE_{VAMP} was significantly more than for $\text{AE}_{\text{K–MEANS}}$. Taking results per pair, we performed binomial significance tests for the 38 pairs of both $\text{AE}_{\text{K–MEANS}}$ and VAE_{VAMP}, followed by Holm-Bonferroni correction over all 76 pairs. After the correction, 10 pairs for $\text{AE}_{\text{K–MEANS}}$ and 16 pairs for VAE_{VAMP} showed significant perceptual difference (corrected $p < 0.005$).

[2]Code is available at github.com/ZackHodari/discrete_intonation

[3]Equivalent to step-wise hard monotonic attention [36, 37] in a sequence-to-sequence model. In the future we'll use an encoder with attention to utilise the learned prosodic features of these models [38].

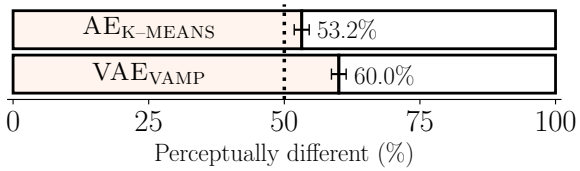

Figure 3: *Same/different results. Error bars shows binomial confidence interval.*

6.2. Qualitative evaluation

While we have shown that VAE_{VAMP} produces distinct renditions more frequently, the above evaluation does not reveal *how* they differ, or what the induced *intonation codes* have captured. To understand this, we first need to know whether our distinct renditions are interpreted differently and if so, how.

Ideally, listeners would identify prosodic constructions linked to specific interpretations, and we would then examine their distributions across *intonation codes*. However, limiting this to previously identified categories or constructions (e.g., [41, 33]) risks missing important types of variation captured by the model. Similarly, a narrow focus on specific linguistic or affective phenomena also increases difficulty for non-expert listeners, and potentially introduces bias. So, to explore this space, we carried out a small, qualitative study. More specifically, we explored: **(i)** whether the prosodic differences captured discourse/information structural or affective differences in meaning, **(ii)** whether *intonation codes* were interpreted in a consistent way across sentences, and **(iii)** what types of variation in prosodic meaning are salient to non-expert listeners.

We took the 6 VAE_{VAMP} pairs with the largest percentage of "different intonation" judgements in the previous test; the mean across listeners for each of these pairs ranged from 76.5% to 82.6%. We ran 45 minute one-on-one interviews with 5 native English-speaking participants (paid £8). We exclude this first (pilot) interview results from the following analysis. We asked listeners to comment on how the sentence was performed and, what effect it had—e.g. did the meaning or emotion change? During the interview, listeners were given all 12 sentences for one *code* pair at a time (i.e. 2 renditions for each of the 12 sentences), and were able to choose which renditions to comment on. Some chose to compare two renditions of a sentence, while others discussed individual renditions independently.

(i) We summarised the interview transcriptions by categorising comments according to descriptive terms; out of 68 total terms, 26 were used to describe more than one sentence. The terms used to describe 4 or more sentences (in order of frequency) were: upset, statement, narrative, question, surprised, "standard" style, continuation rise,[4] emotional, anticipatory, sad, child storytelling, monotonous, and confused. The broad range of terms used is testament to the variety of prosodies our *intonation codes* have captured. However, most of the terms related to more affect-related changes, which is not so surprising given our data. Changes in interpretation relating to information/discourse structure were reported, most notably continuation rise. However, many other stance/interaction related descriptions were also given, e.g. back-channelling, insincere apology/impressed/surprise; and humorous/typical sarcasm.

(ii) Certain *intonation codes* were consistently reported to produce styles such as: questioning, upset, and narrative. In some cases a style was described, but noted as inappropriate (most notably, questioning). Nonetheless, *codes* were not wholly consistent, with their interpretation often changing depending on the sentence. Table 1 shows the number of unique terms, and the terms used multiple times for each sentence. This demonstrates that, unsurprisingly, semantics has a large impact on the perceived effect of the *codes*. The least descriptive sentences, such as "What's the matter now?" and "I'm sorry", elicited the most comments from listeners. This is either because our *intonation codes* are able to produce more variation more freely, or because listeners can imagine more contexts for them. In order to fully determine if individual *codes* behave consistently, we would need a larger sample, and should design sentences specifically for the test.

[4]This term was not used directly, but listeners described the effect.

Table 1: *Single-phrase test sentences: the total number of unique terms used to describe each sentence, and lists of terms used more than once for each sentence.*

13	**There was no answer.** — statement, upset, surprised, anticipatory
11	**"I'm so hungry."** — upset, statement, continuation rise
15	**"Too hard!"** — question, statement
10	**They climbed the stairs.** — upset, continuation rise, anticipatory, sad, narrative
20	**"What's the matter now?"** — statement, question, rhetorical, annoyed, friendly, urgent
11	**"We'd better make sure."** — upset, question, "standard" style, uncertain
12	**"Do you think we're so stupid?"** — insulted, upset, rhetorical, sad
19	**"I'm sorry."** — fake apology, passive aggressive, question, apology, "standard" style, upset
9	**He wanted a turnip.** — statement, narrative, continuation rise, sad, bored
7	**They both tugged and tugged.** — narrative, upset, child storytelling, "standard" style
11	**But the turnip didn't move.** — upset, statement, narrative, surprised
14	**"It's enormous!" cried Jack.** — surprised, exclamation, childlike

(iii) In general, interpretations often appeared dependent on what contexts listeners thought were appropriate for a specific rendition. In fact, some listeners provided rich descriptions of contexts a rendition might make sense in. This could be a useful direction for analysing what a learned representation captures. We could conduct one-on-one interview where listeners are asked to describe some context a rendition might make sense in—selecting "unsure" or "invalid" when necessary. From this descriptive task we could categorise interpretation of different renditions and determine if renditions consistently correspond to plausible, and potentially uncommon, contexts.

Interestingly, users perceived some duration and loudness changes, though neither of these features were modified.

7. Conclusion

We presented a discrete prosodic representation that operates in the phrase domain and produces multiple perceptually distinct renditions of individual sentences. We observed a broad range of affective, and some information structural variation. The interpretation of renditions varied based on semantics, where ambiguity lead to users inventing contexts based on what they perceived. This lead to a new idea for better evaluating the perceived effect of different prosodic renditions, using an interview-based descriptive task.

Acknowledgements: Zack Hodari was supported by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. We thank Oliver Watts for his support both with prosodic phrasing and advice when training VAE_{VAMP}.

8. References

[1] Z. Hodari, O. Watts, and S. King, "Using generative modelling to produce varied intonation for speech synthesis," in *Proc. Speech Synthesis Workshop*, Vienna, Austria, 2019, pp. 239–244.

[2] M. Farrús, C. Lai, and J. D. Moore, "Paragraph-based prosodic cues for speech synthesis applications," in *Proc. Speech Prosody*, Boston, MA, USA, 2016, pp. 1143–1147.

[3] J. Cole and U. Reichel, "What entrainment reveals about the cognitive encoding of prosody and its relation to discourse function," in *Proc. Speech Prosody*, 2016.

[4] J. Kleinhans, M. Farrús, A. Gravano, J. M. Pérez, C. Lai, and L. Wanner, "Using prosody to classify discourse relations," in *Proc. Interspeech*, 2017, pp. 3201–3205.

[5] S. Calhoun, "The centrality of metrical structure in signaling information structure: A probabilistic perspective," *Language*, vol. 86, no. 1, pp. 1–42, 2010.

[6] C. Lai, "Response types and the prosody of declaratives," in *Proc. Speech Prosody*, 2012.

[7] M. E. Armstrong and P. Prieto, "The contribution of context and contour to perceived belief in polar questions," *J. of Pragmatics*, vol. 81, pp. 77–92, 2015.

[8] C. Lai, "What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue," in *Proc. Interspeech*, 2010.

[9] S. Betz, S. Zarrieß, E. Székely, and P. Wagner, "The Greennn Tree — Lengthening Position Influences Uncertainty Perception," in *Proc. Interspeech*, 2019, pp. 3990–3994.

[10] I. Hübscher, M. Garufi, and P. Prieto, "Preschoolers use prosodic mitigation strategies to encode polite stance," in *Proc. Speech Prosody*, 2018, pp. 255–259.

[11] V. Freeman, "Prosodic features of stances in conversation," *J. of the Association for Laboratory Phonology*, vol. 10, no. 1, 2019.

[12] N. G. Ward, J. C. Carlson, and O. Fuentes, "Inferring stance in news broadcasts from prosodic-feature configurations," *Computer Speech and Language*, vol. 50, pp. 85–104, 2018.

[13] J. Yamagishi, T. Masuko, and T. Kobayashi, "HMM-based expressive speech synthesis-towards TTS with arbitrary speaking styles and emotions," in *Proc. of Special Workshop in Maui (SWIM)*, 2004.

[14] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *arXiv preprint arXiv:1807.11470*, 2018.

[15] M. Grice, S. Ritter, H. Niemann, and T. B. Roettger, "Integrating the discreteness and continuity of intonational categories," *J. of Phonetics*, vol. 64, pp. 90–107, 2017.

[16] J. Cole, T. Mahrt, and J. Roy, "Crowd-sourcing prosodic annotation," *Computer Speech and Language*, vol. 45, pp. 300–325, 2017.

[17] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2217–2221.

[18] V. Wan, C. Chan, T. Kenter, J. Vit, and R. Clark, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *Proc. ICML*, Long Beach, USA, 2019.

[19] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[20] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," *arXiv preprint arXiv:1808.01410*, 2018.

[21] S. Tyagi, M. Nicolis, J. Rohnke, T. Drugman, and J. Lorenzo-Trueba, "Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection," *arXiv preprint arXiv:1912.00955*, 2019.

[22] S. Ronanki, G. E. Henter, Z. Wu, and S. King, "A template-based approach for speech synthesis intonation generation using LSTMs," in *Proc. Interspeech*, San Francisco, USA, 2016, pp. 2463–2467.

[23] X. Wang, S. Takaki, J. Yamagishi, S. King, and K. Tokuda, "A vector quantized variational autoencoder (vq-vae) autoregressive neural f0 model for statistical parametric speech synthesis," *IEEE Trans. on Audio, Speech and Language Processing*, 2019.

[24] J. Cole, T. Mahrt, and J. I. Hualde, "Listening for sound, listening for meaning: Task effects on prosodic transcription," in *Proc. Speech Prosody*, 2014, pp. 859–863.

[25] R. Turnbull, A. J. Royer, K. Ito, and S. R. Speer, "Prominence perception is dependent on phonology, semantics, and awareness of discourse," *Language, Cognition and Neuroscience*, vol. 32, no. 8, pp. 1017–1033, 2017.

[26] A. Köhn, T. Baumann, and O. Dörfler, "An empirical analysis of the correlation of syntax and prosody," in *Proc. Interspeech*, 2018, pp. 2157–2161.

[27] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008, ch. 8.

[28] M. Y. Liberman and K. W. Church, "Text analysis and word pronunciation in text-to-speech synthesis," *Furui and Sondhi, Advances in Speech Technology*, pp. 791–832, 1992.

[29] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[30] J. M. Tomczak and M. Welling, "VAE with a VampPrior," in *Proc. Artificial Intelligence and Statistics*, Lanzarote, Spain, 2018, pp. 1214–1223.

[31] S. King, J. Crumlish, A. Martin, and L. Wihlborg, "The Blizzard challenge 2018," in *Proc. Blizzard Challenge Workshop*, Hyderabad, India, 2017.

[32] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector," in *Proc. ICASSP*. IEEE, 2019, pp. 6925–6929.

[33] D. Goodhue, L. Harrison, Y. C. Su, and M. Wagner, "Toward a bestiary of English intonational contours," in *Proc. North East Linguistics Society*, 2016, pp. 311–320.

[34] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3. Istanbul, Turkey: IEEE, 2000, pp. 1315–1318.

[35] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[36] Y. Yasuda, X. Wang, and J. Yamagishi, "Initial investigation of encoder-decoder end-to-end TTS using marginalization of monotonic hard alignments," in *Proc. Speech Synthesis Workshop*, 2019, pp. 211–216.

[37] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS," in *Proc. Interspeech*, 2019, pp. 1293–1297.

[38] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data," in *Proc. Interspeech*, 2019.

[39] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, Long Beach, USA, Dec 2017, pp. 5998–6008.

[41] N. G. Ward, *Prosodic Patterns in English Conversation*. Cambridge University Press, 2019.